

Data matters: how earth and environmental scientists determine data relevance and reusability

Angela P. Murillo

School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, USA

Abstract

Purpose – The purpose of this study is to examine the information needs of earth and environmental scientists regarding how they determine data reusability and relevance. Additionally, this study provides strategies for the development of data collections and recommendations for data management and curation for information professionals working alongside researchers.

Design/methodology/approach – This study uses a multi-phase mixed-method approach. The test environment is the DataONE data repository. Phase 1 includes a qualitative and quantitative content analysis of deposited data. Phase 2 consists of a quasi-experiment think-aloud study. This paper reports mainly on Phase 2.

Findings – This study identifies earth and environmental scientists' information needs to determine data reusability. The findings include a need for information regarding research methods, instruments and data descriptions when determining data reusability, as well as a restructuring of data abstracts. Additional findings include reorganizing of the data record layout and data citation information.

Research limitations/implications – While this study was limited to earth and environmental science data, the findings provide feedback for scientists in other disciplines, as earth and environmental science is a highly interdisciplinary scientific domain that pulls from many disciplines, including biology, ecology and geology, and additionally there has been a significant increase in interdisciplinary research in many scientific fields.

Practical implications – The practical implications include concrete feedback to data librarians, data curators and repository managers, as well as other information professionals as to the information needs of scientists reusing data. The suggestions could be implemented to improve consultative practices when working alongside scientists regarding data deposition and data creation. These suggestions could improve policies for data repositories through direct feedback from scientists. These suggestions could be implemented to improve how data repositories are created and what should be considered mandatory information and secondary information to improve the reusability of data.

Social implications – By examining the information needs of earth and environmental scientists reusing data, this study provides feedback that could change current practices in data deposition, which ultimately could improve the potentiality of data reuse.

Originality/value – While there has been research conducted on data sharing and reuse, this study provides more detailed granularity regarding what information is needed to determine reusability. This study sets itself apart by not focusing on social motivators and demotivators, but by focusing on information provided in a data record.

Keywords Data reuse, Research data management, Data sharing, Data curation, Data repositories, Scientific data

Paper type Research paper

Introduction

While much effort has been taken to create data repositories for sharing and reuse, there has been less attention to examining what is necessary for these data to be successfully reused. Efforts in the creation of data repositories, the creation of data sharing policies and tools to assist with sharing and reuse have propelled the ability for scientists to share and reuse data. Additionally, research to examine motivations and inhibitors for data sharing and reuse have been well-documented. While this research has been instrumental in understanding what motivates scientists to share and reuse data, an under-researched area of study is to

consider what scientists need to know about data sets to determine reusability as scientist have a vast amount of available data to reuse from the many data repositories in existence.

When considering data reusability, it is essential to consider that “reusability can be only appraised from the potential reuser perspective, who will juxtapose best judgment about the attributes of the available data to their reuse intention/purpose” (Yoon *et al.*, 2017, p. 2). This study considers that the potential reuser is provided a variety of attributes about a data set from a data records, and from that data record determines reusability and relevance of a data set. Faniel and Jacobsen (2010) describe three considerations when assessing data for reusability:

- 1 Are the data relevant?
- 2 Can the data be understood?
- 3 Are the data trustworthy?

The current issue and full text archive of this journal is available on Emerald Insight at: www.emeraldinsight.com/2514-9326.htm



Collection and Curation
© Emerald Publishing Limited [ISSN 2514-9326]
[DOI 10.1108/CC-11-2018-0023]

Received 20 November 2018
Revised 8 January 2019
21 February 2019
20 March 2019
Accepted 1 May 2019

This study focuses on Points 1 and 2 by examining how scientists determine relevance and if scientists can understand the data through the data record provided. A data record can be defined as structured information that present essential information of their host page, product, service (Liu, Grossman and Zhai, 2003), or other items they are representing, in this case, data. Data reuse has been defined as secondary use of data other than originally intended (Faniel and Jacobsen, 2010; Zimmerman, 2008).

This study addresses the questions: *How do scientists determine data reusability and relevance*, and more specifically, *what information or attributes about the data do scientists need to determine data reusability and relevance?* This paper explores the research questions through the results of a multi-phased mixed-methods study. The first phase of the study examines the attributes provided in data records through a qualitative and quantitative content analysis of data records. The second phase of this study examines which attributes assist scientists in determining data reusability and relevance through a quasi-experiment think-aloud study.

The findings of this study provide librarians, curators, and repository managers a better understanding how earth and environmental scientists determine data reusability and relevance, and provides recommendations for creating data records with the greatest potential for reuse.

Literature review

Data sharing and reuse have been extensively studied and well documented. Data sharing and reuse has many benefits including the ability to extract additional value from data, enable reproducible research, enable others to ask new questions of existing data, and advance science (Borgman, 2010, 2012; Lord and Macdonald, 2003).

Changes in policy have created an environment where scientists are encouraged and sometimes required to share data through data sharing policies from grant funding agencies such as the National Institutes of Health (National Institutes of Health, 2003, 2007) and National Science Foundation (National Science Foundation, 2010). Moreover, journals are more frequently encouraging or requiring the sharing of scientific data for publication of research (Brown, 2003; McCain, 1995).

As the policies promoting data sharing and reuse became more prevalent, there has also been an increase in the availability of scientific data repositories (Marcial and Hemminger, 2010) for scientists to share and reuse data. Moreover, an increase of integrated systems for data sharing and discovery such as USGIN (U.S. Geoscience Information Network, n.d.), Dryad (2019), and DataONE (DataONE, 2013b) have become available for scientists.

Both the changes in policies and technology have increased the research and examination of factors impacting data sharing and reuse. Motivations and inhibitors for data sharing and reuse have additionally been thoroughly researched. Previous literature has focused on motivators such as scientific reputation (Ceci, 1988; Sieber, 1988) and the value of data and data duplication (Borgman, 2012; Lord and Macdonald, 2003). Additionally, literature has focused on inhibitors including financial concerns, time for reusing data, and effort in

creating data (Cohen, 1995; Tenopir et al., 2011). Studies have also focused on various types of data withholding (Blumenthal et al., 2006; Noor et al., 2006) and views of data ownership (Constant et al., 1994). Additionally, there has been an exploration of potential data reusers (Zimmerman, 2008). Tenopir et al. (2015) examined change over time of data sharing and reuse practices, and perceptions of organizational support for data sharing and reuse. Si et al. (2015) evaluated scientific data sharing platforms for performance evaluations, focusing on operation management, data resource, platform function, and efficiency; however, these evaluations were not conducted by scientists testing the system. Fecher et al. (2015) developed a conceptual framework of data sharing from the researcher's perspective. Additionally, researchers have examined how communication impacts data reusers (Yoon, 2017). Most closely related to this study, Joo and Kim's (2017) examined data reuse behaviors of engineering researchers, however, still focused on attitudes towards data reuse.

While this previous research is essential in understanding data sharing and reuse there is a thematic and methodological gap in the literature. Thematically, as noted previously, much of this research focuses on incentive, disincentives, and attitudes towards data sharing and reuse, as well as policy supporting data sharing and reuse. Methodologically much of this research has been conducted through self-reporting measures such as interviews and surveys (Blumenthal et al., 2006; Sayogo and Pardo, 2013; Tenopir et al., 2015). Additionally, there have been bibliometric studies of data deposition and data citation (Piwowar, 2011; Piwowar and Chapman, 2010). There are very few experimental studies (Constant et al., 1994) and a lack of mix-methods studies which can provide a richer understanding of how scientists determine data reusability and relevance while searching for data in data repositories.

Ultimately how scientists determine the reusability of shared data and what scientists need to know about these data to deem them appropriate for reuse is understudied. This study explores this topic through a multi-phase mixed-method approach.

Methods

This study uses a multi-phase mixed-method approach including a qualitative and quantitative content analysis (Phase 1) and a quasi-experimental think-aloud study (Phase 2). While this paper focuses on the findings of the think-aloud quasi-experiment study, it is important to include a description of the content analysis conducted in Phase 1 to understand the study design.

DataONE was chosen as the test environment for several reasons. DataONE provides the ability for scientists to share and reuse data within the DataONE system, thus providing the ability to examine data sharing and data reuse within the same environment. While DataONE focuses on the earth and environmental sciences, these sciences are particularly interesting for examining data sharing and reuse because of their interdisciplinary nature. The earth and environmental sciences are highly interdisciplinary fields with many subfields, data types, and methods of data collection. As science is becoming more interdisciplinary and transdisciplinary (Baker, 2015), examining data sharing and reuse within an already interdisciplinary field with heterogeneous data allows for

potentially more generalizable results as the growth in interdisciplinary and transdisciplinary research has greatly increased and are vital to addressing complex scientific challenges (Hall *et al.*, 2018).

The DataONE data repository allows scientists to search over 800,000 data objects through a free online search interface which searches a federation of repositories (DataONE, 2013b). The search interface includes a simple search with facets for search refining, familiar to most researchers. Figure 1 shows an example of a data record provided by DataONE. While the entire data record is not included for space, Figure 1 provides an understanding of the types of information provided in the data record. As shown in Figure 1, a DataONE data record includes information about the data such as a data identifier, data abstract, keywords, funding information, research methods, and sampling. This information is provided to scientists to assist in determining data reusability.

This study was interested in how scientists determine data reusability and what factors impact data relevance. Therefore, this study needed to be conducted in two phases. Phase 1 examines what information is shared regarding the data through a qualitative and quantitative content analysis of data records. Phase 2 examines what scientist need to determine reusability through a quasi-experiment think-aloud study. Two pilot studies were conducted to test, evaluate, and refine the research methods of both phases.

The first pilot test examined data shared within the DataONE repository. A random sample of 650 data records

extracted from the DataONE to test the sampling method and consider qualitative and quantitative variables associated with the data. From this study, it was determined that a random sample produced an overrepresentation of data from certain Member Nodes. DataONE Member Nodes are data repositories that expose their data and metadata to the DataONE (DataONE, 2013a). This overrepresentation led to a stratified sample for the Phase 1 final sampling.

The second pilot study (Murillo, 2014) explored the usefulness of a think-aloud approach for examining data reuse. For this study, six users searched the DataONE for data to reuse. Participants were asked to think-aloud and describe how they determine reusability. From this study, it was determined that the think-aloud approach produced too many uncontrollable factors such as user-designed search queries and too many data record results, which led to the final quasi-experiment design.

Phase 1: research methods

Phase 1 examined information provided about the data through quantitative and qualitative content analysis of 202 data records. As the purpose of this phase of the study was to gain an understanding of the data shared within the DataONE, exploratory descriptive content analysis was used to summarize the records, as it focuses on the features of recorded information (Spurgin and Wildemuth, 2009, p. 298). A combination of inductive and deductive analysis was used as is common in exploratory content analysis (Neuendorf, 2002, pp. 11-12).

Figure 1 DataONE search result data record

General	
Identifier	doi:10.5063/F19K48G6
Abstract	Stream temperature is an important parameter to ecology, climate, and hydrology studies in Alaska. The National Park Service (NPS) has collected hourly stream temperature (C) data throughout southwest Alaska. Sampling for each site occurred at some point between 2006 to 2017; each sampling site has at least 1 year of data and at most 10 years. Most of the stream temperature monitoring sites are ongoing; however, some sites are missing data from 2015-2016 due to losing the loggers. One site (LakeClark, 1935, 2014, 2016.csv) is missing data from 2016-2017 due to being unable to download the logger during the 2017 field season. Some sites have data taken from lakes and have multiple depths, anywhere from 0-240 m. This dataset is part of a larger project to collect a comprehensive statewide inventory of current and historic continuous monitoring locations for stream and lake temperatures. These data were provided by NPS for archival as part of an effort between the State of Alaska Salmon and People (SASAP, https://alaskasalmonandpeople.org/) and the Alaska Center for Conservation Science's Alaska Online Aquatic Temperature Site (AKOATS, http://aocs.uaa.alaska.edu/aquatic-ecology/akoats/) to make stream temperature data more readily available for researchers. This package includes 17 stream temperature data files. File names are formatted WaterbodyName_AKOATSID_StartDate_EndDate. This package also includes site level metadata (SiteLevelMetadata_Bartz.csv).
Keywords	Keyword
	Type
	stream
	temperature
	southwest
	stream temperature
	southwest Alaska
	NPS
National Park Service	
Funding:	
State of Alaska's Salmon and People (Gordon and Betty Moore Foundation Award 5124)	
Data Task Forces for Better Synthesis Studies (Gordon and Betty Moore Foundation Award 5451)	
Methods	Step 1
	Description
	Stream temperature sampling
	Stream temperature (C) was sampled at 17 sites throughout the Southwest Alaska Network using dataloggers. Data was logged at regular intervals of 60 minutes. Based on regional, multi-agency discussions on temperature monitoring data collection standards, there is a minimum accuracy standard of +/- 0.25°C.
Step 2	Description
	Data quality assurance
	Some data were QC'ed by NCEAS; each site's data was plotted (temperature over time), and obvious outliers, data that measured air temperatures, and constant readings of 0 degrees Celsius indicating the logger was frozen in ice over the winter were given a '0' (false) Boolean value in the UseData attribute. The data QC'ed prior to NCEAS obtaining data were only given '0' for missing sampleDate and sampleTime information.
Sampling	
Sampling Step 1	
Sampling Area And Frequency	Sampling occurred at some point between 2006 to 2017; each sampling unit has at least 1 year of data and at most 10 years. Not all sites logged data for every year in the study period.
Sampling Description	Sites are selected to be stable locations within the reach in a well-mixed section of the stream channel. Care is taken to place the logger in the active channel to prevent risk of exposure to air during low flows.

A stratified sample of ten records from each Member Node of the DataONE was used for the sampling frame. As discussed, the Phase 1 pilot study indicated that a random sample did not accurately represent the data records available in the DataONE, as Member Nodes with more data were overrepresented.

The content analysis steps included the conceptualization of variables through the pilot study, the operationalization of the variables, coding and training, and final coding (Neuendorf, 2002). Two coders independently, both with previous experience coding qualitative and quantitative data, coded a subset of the data, created a codebook, and then coded the entire data set. Final intercoder reliability was 0.91 Krippendorff's alpha. Krippendorff's alpha adjusts for whether the variable is measured as nominal, ordinal, interval, or ratio and is a highly attractive coefficient but rarely used because of the difficulty in calculating (Neuendorf, 2002, p. 151). In the case of this study, this alpha was chosen to compute intercoder reliability because of the variability in the variables, which included nominal and ordinal variables.

Phase 2: research methods

Phase 1 provided the details of the data records. Phase 2 used these detailed data records to create an experimental interface for the quasi-experiment think-aloud study.

The think-aloud method can be helpful to understanding decision-making and knowledge of a system, and participants have little to no memory or interpretation errors, and while there may be some cognitive disturbances, these are minimal (Somerén *et al.*, 1994). The quasi-experiment counter-balance design is typically used to test search interfaces by applying multiple treatments to each participant (Hank & Wildemuth, 2009), and in this case, was used to test the relevance of data records.

An experimental interface used a counter-balanced design where each participant rotated through manipulated data records based on a pre-defined query. From the pilot study, it became clear that the search query and search results needed to be predefined to control the test environment.

The query "soil moisture content" was used and four manipulated results were created. This query and results were used because it was broad enough to be applicable to many sciences and the results were robust enough to be able to develop several versions of results for the quasi-experiment. Additionally, feedback received from scientists during the pilot study verified the rationale regarding the pre-defined search query and results.

The four manipulated results were created based on the findings of Phase 1. Result 1 contained all 27 unique pieces of information or attributes found in the data records including an abstract, research methods section, unit and attribute list, a data description, and all attributes described in the Phase 1 Findings. Result 2 contained all attributes except the unit and attribute list. Result 3 contained all attributes except the research methods, unit, and attribute list. And lastly, Result 4 contained all attributes except the abstract, unit, and attribute list.

The results were manipulated through a counterbalanced design, and the results were rotated with each participant, as shown in Table I.

Table I Counterbalanced design

Participant (P)	Query result			
P1	Result 1	Result 2	Result 3	Result 4
P2	Result 2	Result 4	Result 1	Result 3
P3	Result 3	Result 1	Result 4	Result 2
P4	Result 4	Result 3	Result 2	Result 1

Figure 2 provides an overview of the full procedures of the quasi-experiment think-aloud study design and indicates the points and types of data collection throughout. Each participant was first provided an online consent form. Secondly, participants were provided with a sample data record to practice "thinking-aloud" about what information assisted them in determining data reusability. Next, participants were provided four data records and asked to think aloud for each, after each result they were provided a post-result usefulness survey (Appendix 1). Once participants completed thinking-aloud for each result, they were asked to complete a rank-order survey (Appendix 2). Next, participants were asked to complete a post-experiment survey (Appendix 3), which included open-ended questions, data reuse factors questions, and demographic questions. Lastly, a semi-structured interview was conducted to have participants elaborate on their current data reuse practices, as well as how they determine data reusability.

Scientists were recruited through the University of North Carolina and North Carolina State University departmental listservs for the geological sciences, environmental sciences, ecology, and biology. Additionally, participants were recruited through the Committee on Data for Science and Technology (CODATA) and DataONE listserv. Lastly, scientists were recruited at the Annual Geological Society of America annual conference.

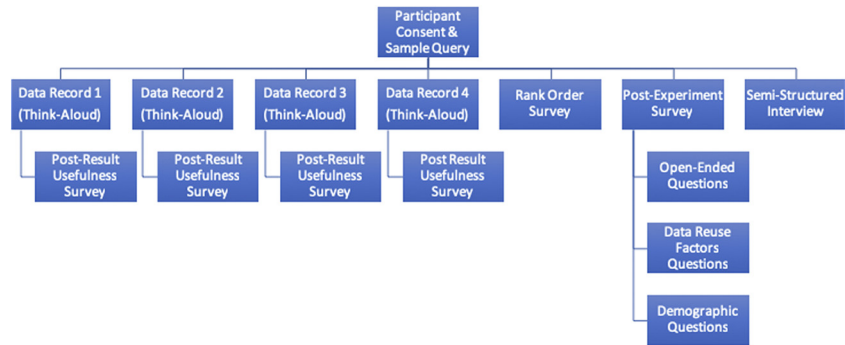
A total of 16 participants were recruited. Scientists were asked to "think aloud" regarding how they determine data reusability. Additionally, participants submitted a post-result usefulness survey after each manipulated result, and a post-experiment data reuse factors survey. Lastly, semi-structured interviews were conducted after the experiment.

The data from the quasi-experiment and surveys were analyzed using descriptive statistics. The think-aloud notes and the semi-structured interview notes were analyzed using inductive content analysis.

Findings

Phase 1: qualitative and quantitative content analysis

The examination of the data records provided 27 unique pieces of information or attributes regarding the data including: Data Citation, Instrument, Geographic, Intellectual Rights, Dataset Metadata, Research Methods, Funding Source, Data Availability, Access Metadata, Taxonomic, Abstract, File Size, Metadata Standard, Publisher, Additional Access, Data Type, Additional Metadata Standard, Attribute List, Keywords, Creator, Provenance, Unit List, Keyword Thesauri, Publication Date, Temporal, Contact, Associated Party and Data Description. Not every data record contained all pieces of information.

Figure 2 Quasi-experiment think-aloud study procedures

The records varied when it came to the information provided. Only 51 per cent contained information about the majority of the items listed above. Items that were particularly well covered were creator (98.1 per cent), keywords (98 per cent) and abstract (87 per cent). Items that were particularly not well covered included funding source (17 per cent), taxonomic (14 per cent) and data description (12 per cent).

As discussed previously, the findings from Phase 1 were used to build the manipulated data record results for the quasi-experiment think-aloud study. The decision on which attributes to include in the results was based on the findings from Phase 1, as Phase 1 determined what percentage of each attribute was included in the data records. Each manipulated result emulated the range of data records found. Result 1 represented the most robust results found in Phase 1, and each subsequent result removed the next most likely attributes not be included in the data record. The manipulated results also received feedback from scientists during pilot testing of the experimental interface.

Phase 2: quasi-experiment think-aloud, post-result usefulness survey and post-experiment data reuse factors survey

Of the 16 participants, 56 per cent (9) were male and 44 per cent (7) were female. Six participants were geologists, four were ecologists, two were atmospheric scientists, two were environmental scientists, one was a physicist, and one was a hydrologist. Additionally, 31.3 per cent had PhDs, 37.5 per cent had master's degrees, and 31.3 per cent had bachelor's degrees. None of the participants had previously used DataONE. The average time spent through the quasi-experiment think-aloud, post-experiment survey, and semi-structure interview was approximately 90 min.

From the post-result usefulness survey (Appendix 1) overwhelmingly, the scientists found Result 1 to be the most useful (81 per cent) when it came to determining data reuse.

This particular result included very robust information and included an abstract, research methods, attribute lists, instrument information, as well as all 27 attributes describes in the Phase 1 Findings. While overwhelmingly Result 3 was seen as the least useful (81 per cent). Result 2 was considered the second most useful. Result 4 was considered the third most useful. The findings of the usefulness survey indicated a proclivity to more robust research methods information, data description and unit/attribute information tended to yield more useful results. Table II below summarizes the Post-result usefulness survey.

The rank-order survey (Appendix 2), which was administered after participants though-aloud through all of the results provided similar results to the post-result usefulness survey. Participants were asked to rank all results in the order of most useful to least useful in regard to assisting their ability to reuse the data. This survey provided the same results as the post-result usefulness survey, therefore even after the participants interacted with each result, the perceived Result 1 as the most useful and Result 3 as the least useful.

The data reuse questions in the post-experiment survey results (Appendix 3, Table III) show that scientists found the attribute list, data description, and research methods information particularly important when determining data

Table III Data reuse questions results ($N = 16$)

Data reuse factors (scale 1-7)	Mean (SD)
Attribute list	6.60 (0.52)
Data description	6.50 (0.49)
Research methods information	6.13 (0.49)
Instrument information	5.88 (0.52)
Provenance information	5.25 (1.18)
Metadata standard	4.94 (1.53)
Intellectual property information	4.75 (1.29)

Table II Post-result usefulness survey summary

Most useful		Least useful	
Result 1 (3.56 mean)	Result 2 (3.31 mean)	Result 4 (2.31 mean)	Result 3 (2.25 mean)
All attributes	All attributes except unit and attribute list	All attributes except abstract, unit, and attribute list	All attributes except research methods, unit, and attribute list

reusability. Additionally, these results indicated that metadata standard and intellectual property information was not particularly important in determining data reuse. Participants were required to rank every factor in this portion of the post-experiment survey.

Qualitative results

Qualitative analysis of the think-aloud and semi-structured interviews provided further detail regarding how scientists determine data reusability. Scientists tend to rely on the data description, research methods information, unit, and attribute lists to determine data reusability. Scientists relied less on the abstract information to determine reusability.

Additionally, participants found the data description was the most pertinent information. The data description is a short and succinct summary of the data set presented near the bottom of the result. Participants suggested that this vital piece of information moved to the top of the data record. Participants stated that the organization of the information should be considered on a data record. Participants considered this primary information to be the data description, research methods information, format, size, and data type, and attribute/unit lists. Participants considered secondary information to be keywords, metadata standard and intellectual property information.

Participants discussed why Results 3 and 4 were the least useful results. Several participants discussed how both of these results did not provide enough information to determine reusability without downloading. Participants stated that these results were “what they were used to with other data repositories” and that they have in the past wasted time downloading data just to learn that it was not relevant to their data needs.

Furthermore, scientists described the need for basic information including format (.png, .csv), size, and type (experimental, field, sensor). Participants were frustrated that this information was not clear on the data record and see it as basic information regarding data that should be included in all data records. One participant stated how problematic that size and format was not included in the data record by stating:

For instance, if I needed image files in .PNG format, it would save time if I knew that a dataset were .JPG only. Similarly, knowing the size of the dataset could be critical. I don't want to download 100 TBs worth of data just to have my computer crash (P11).

Another participant stated that data repository managers and data sharers should consider the “*who (data creator), what (data type), when (when collected), where (where collected) and how (research methods) of data*” (P5) as primary information that should always be included in data records.

During the semi-structured interviews, it was determined that all of the participants had some previous experience reusing data. Many of them used similar data repositories as the DataONE, or found data associated with publications for data reuse. They described searching for data as an arduous process, as many repositories did not have the extensive amount of information as the data records they had just encountered through the study. The participants stated that their main inhibitor for data reuse was the lack of information about the data.

Discussion

As discussed in the findings, the participants of this study had previous experience in data reuse. Participants described their data reuse experiences as first looking at literature and then acquiring the data through the data owner. This tactic is consistent with the literature (Zimmerman, 2003, 2008). Some participants described using data libraries including NOAA and USGS, and other data repositories available to scientists (Marcial and Hemminger, 2010). However, most of the participants additionally described a need for quality control of data for reuse (Baru, 2007), the need for improved access and discover (Beran *et al.*, 2010), and a lack of time and support to search for data to reuse (Tenopir *et al.*, 2011).

Scientists' information needs regarding data are different than the needs of researchers looking for research articles. The majority of the scientists indicated that abstracts were not useful to their ability to determine data reusability because it was not clear if the abstract were referring to the paper associated with the data or the data itself, and too often the abstract provided information about the project the data was created for. It frustrated scientists to read through long text to realize that the abstract did not provide any pertinent information about the data.

Overwhelmingly scientists suggested that the research methods were crucial for determining reusability. Research methods provided information regarding collection, instruments, and manipulation. One scientist stated that this answered the “who, what, when, where, and how” (P5) about the data. This approach has been suggested in the previous research (Baru, 2007), and additionally has been discussed in data reproducibility literature (Lifschitz *et al.*, 2011). During the post-experiment interviews, the scientists discussed how the research methods provide the most precise description of how the data was created and collected, and therefore provides scientists the most complete understanding of the data itself. Additionally, scientists stated that the research methods allowed them to determine if the data was collected as rigorously as they would want it to be, and allowed them to judge the meticulousness, thoroughness, and appropriateness of the research methods used to collect the data. In this sense, the research methods allow scientists the ability to judge both their understanding and trustworthiness of the data, as described in the literature (Faniel and Jacobsen, 2010).

Scientists also noted that without calibration information of the instruments this could become less useful information. They stated how calibration information, instrumentation information, and provenance information is often left out of data records and this information could be vital to the reusability of data.

Scientists discussed how they appreciated a suggested data citation format and DOI. Nearly all scientists stated that this information would encourage data citation, literature has also indicated this preference (Piwowar and Vision, 2013). Scientists suggested they prefer when the data record contains a simple suggested data citation format that they can easily copy and paste into their own publication.

Scientists were surprised that the keywords were not linked throughout the system, as they were used to being able to click on keywords to link to similar results in most systems. It would

be worth considering adding this and other types of recommendations to data repositories, similar to what most scientists are used to while looking for research articles.

Information overload was discussed, and surprisingly, most scientists suggested that they prefer too much information than not enough information. This was counter to what much of the research indicates regarding information fatigue (Eppler and Mengis, 2004). However, some literature suggests that search stopping behavior is dependent on task (Browne *et al.*, 2007). In the case of data reuse, the participants of this study stated they would rather not want to waste time downloading data to find out it did not meet their needs. In addition to this, scientists also did not mind that all of the information was on one page and that they had to scroll, however, they did suggest that drop-down menus could assist with long data records. Additionally, scientists suggested that including mouseover definitions of data attributes would be helpful.

Recommendations for information professionals

When information professionals are working with scientists to assist with depositing data into repositories or providing feedback regarding data management plans, the following recommendations should be considered to assist data creators in producing reusable data:

- detailed research methods;
- including calibration information with instrument information;
- data format, data size, and data type;
- data citation format and DOI;
- succinct description of the data;
- considering what is primary information versus secondary information when creating the layout in data records;
- considering drop-down menus to organize long data records; and
- linking keywords throughout the system.

Data record prototype

Updating the information in the data record, as well as the organization of this information, could improve the potential reusability of data and would assist potential data reusers to determine relevance more easily. A preliminary prototype of an ideal data record is provided in Table IV which synthesizes the findings of the quasi-experiment think-aloud results, as well as the qualitative data from the semi-structured interviews following the experiment. Table IV is divided into primary information and secondary information. Scientists suggested that specific data attributes were more critical for them to determine data reusability (primary information), while other information they appreciated having but did not necessarily impact their ability to assess reusability (secondary information).

Table IV provides a set of primary and secondary attributes that scientists have suggested assist in their determination of data relevance and reusability and can be used in future studies of data reuse and data records, as well as provide guidance for information professionals working with data collections.

Conclusion

While this study provides a more thorough understanding of how earth environmental scientists determine data reusability,

Table IV Ideal data record attributes and definitions

Attribute	Attribute definition
Primary information	
Data description	Short and succinct data description
Data creator	Who collected the data?
Data format	What is the format of the data? (.csv, .txt, .tif, etc.)
Data type	What type of data was collected? (field, experiment, sensor, simulation)
Data size	What is the size of the data? (MBs, TBs, etc.)
Data collection location	Where was the data collected?
Data date range	When was the data collected?
Research methods information	How was the data collected, by what means, what were the steps involved?
Instrument information	What instruments were used to collect the data and what were the calibration settings?
Provenance information	Was the data changed in any way, if so how, why, and by what methods and/or instrument?
Data abstract	A descriptive summary of the data. The data abstract should describe the data, not the paper associated with the data
Attribute and unit lists	This includes data variables and how these were measured
Secondary information	
Taxonomic information	If appropriate for the data set, any information regarding biological organisms
Data citation and persistent identifier	A suggested data citation format, and DOI or other persistent identifier for the data
Intellectual rights information	Any statement regarding restrictions to use of the data, as well as attribution instructions
Data keywords	Keywords linked throughout repository so that potential reusers can click to similar data sets
Metadata standard	Metadata standard used by the data
Funding source	Funding source for the data collection
Publication date	Date the data was published

there are still limitations to the study. This study examined one single data repository (DataONE) in one discipline (earth and environmental science). While earth and environmental science is a highly interdisciplinary field that contains many subdisciplines, the findings in this study do not represent all scientific disciplines. Another limitation to this study was the use of a hypothetical search for the quasi-experiment think-aloud. While the pilot studies indicated a need to control this search results, natural searches from the participants may have provided other insights for how scientists determine data reusability. Lastly, the sample size of the quasi-experiment think-aloud is rather small. While major recruitment efforts were made, additional participants could have potentially provided more information. However, even with the small sample size, there was a fairly clear consensus amongst the 16

participants. Additionally, having 16 participants allowed for four rotations of each result in the counterbalanced design of the quasi-experiment.

This study contributes to a greater understanding of how scientists determine data reusability through a quasi-experiment think-aloud study. This study provides new contributions to the current research in several ways. From a research perspective, much of the previous research in data reuse focuses on how policy influences data reuse, as well as motivators and inhibitors of data reuse such as time and effort needed to reuse data. While much work has been done in increasing the availability of data for reuse, there is still work that needs to be done to ensure data reusability. Rarely has the previous research examined data repositories and attributes provided about data for reuse through a data record.

These findings have implications for a broader audience including data sharing organizations, research data management organizations, and data management plan creators. For example, these findings can assist creators of data management plan templates and resources such as the DMPTool (University of California Curation Center, 2019) to ensure that they are addressing the specific needs of data reusers.

This study provides direct feedback from scientists about the data record itself and provides a set of recommendations for data creators and information professionals to ensure the greatest potential for future reuse. Additionally, it provides feedback to data repository managers to consider when building data repositories.

References

- Baker, B. (2015), "The science of team science", *BioScience*, Vol. 65 No. 7, pp. 639-644, available at: <https://doi.org/10.1093/biosci/biv077>
- Baru, C. (2007), "Sharing and caring of eScience data", *International Journal on Digital Libraries*, Vol. 7 Nos. 1/2, pp. 113-116, available at: <https://doi.org/10.1007/s00799-007-0029-2>
- Beran, B., van Ingen, C. and Fatland, D.R. (2010), "SciScope: a participatory geoscientific web application", *Concurrency and Computation: Practice and Experience*, Vol. 22 No. 17, pp. 2300-2312, available at: <https://doi.org/10.1002/cpe.1597>
- Blumenthal, D., Campbell, E.G., Gokhale, M., Yucel, R., Clarridge, B. and Hilgartner, S. (2006), "Data withholding in genetics and the other life sciences: prevalences and predictors", *Academic Medicine*, Vol. 81 No. 2, pp. 137-145.
- Borgman, C.L. (2010), "Research data: who will share what, with whom, when, and why?", *Fifth China – North America Library Conference 2010, (September)*, available at: <http://works.bepress.com/borgman/238/>
- Borgman, C.L. (2012), "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078, available at: <https://doi.org/10.1002/asi.22634>
- Brown, C.M. (2003), "The changing face of scientific discourse: analysis of genomic and proteomic database usage and acceptance", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 10, pp. 926-938, available at: <https://doi.org/10.1002/asi.10289>
- Browne, G.L., Pitts, M.G. and Wetherbe, J.C. (2007), "Cognitive stopping rules for terminating information search in online tasks", *MIS Quarterly*, Vol. 31 No. 1, pp. 88-104.
- Ceci, S.J. (1988), "Scientists' attitudes toward data sharing", *Science, Technology, and Human Values*, Vol. 13 Nos 1/2, pp. 45-52.
- Cohen, J. (1995), "Share and share alike isn't always the rule in science", *Science (New York, N.Y.)*, Vol. 268 No. 5218, pp. 1715-1718.
- Constant, D., Kiesler, S. and Sproull, L. (1994), "What's mine is ours, or is it? A study of attitudes about information sharing", *Information Systems Research*, Vol. 5 No. 4, pp. 400-421.
- DataONE (2013a), "Benefits of becoming a member node | DataONE", available at: www.dataone.org/benefits-becoming-member-node (accessed 2 January 2014).
- DataONE (2013b), "What is DataONE? | DataONE", available at: www.dataone.org/what-dataone (accessed 2 January 2014).
- Dryad (2019). "Dryad", available at: <http://datadryad.org/> (accessed 7 October 2011).
- Eppler, M.J. and Mengis, J. (2004), "The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines", *The Information Society*, Vol. 20 No. 5, pp. 325-344, available at: <https://doi.org/10.1080/01972240490507974>
- Faniel, I.M. and Jacobsen, T.E. (2010), "Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data", *Computer Supported Cooperative Work (Cscw)*, Vol. 19 Nos. 3/4, pp. 355-375, available at: <https://doi.org/10.1007/s10606-010-9117-8>
- Fecher, B., Friesike, S. and Hebing, M. (2015), "What drives academic data sharing?", *Plos One*, Vol. 10 No. 2, p. e0118053, available at: <https://doi.org/10.1371/journal.pone.0118053>
- Hall, K.L., Vogel, A.L., Huang, G.C., Serrano, K.J., Rice, E. L., Tsakraklides, S.P. and Fiore, S.M. (2018), "The science of team science: a review of the empirical evidence and research gaps on collaboration in science", *American Psychologist*, Vol. 73 No. 4, pp. 532-548, available at: <https://doi.org/10.1037/amp0000319>
- Hank, C. and Wildemuth, B.M. (2009), "Quasi-experimental studies", in Wildemuth, B. M. (Ed.), *Applications of Social Research Methods to Questions in Information and Library Science*, Libraries Unlimited, Westport, Conn, pp. 93-104.
- Joo, Y.K., and Kim, Y. (2017), "Engineering researchers' data reuse behaviours: a structural equation modelling approach", *The Electronic Library*, Vol. 35 No. 6, pp. 1141-1161, available at: <https://doi.org/10.1108/EL-08-2016-0163>
- Lifschitz, S., Gomes, L. and Rehen, S.K. (2011), "Dealing with reusability and reproducibility for scientific workflows", *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops, IEEE, Atlanta, GA*, pp. 625-632.

- Liu, B., Grossman, R. and Zhai, Y. (2003), "Mining data records in web pages", *Presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 601-606, available at: <https://doi.org/10.1145/956750.956826>
- Lord, P. and Macdonald, A. (2003), *e-Science Curation Report: Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision*, The JISC Committee for the Support of Research (JCSR), Twickenham, England, pp. 1-84, available at: www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
- McCain, K.W. (1995), "Mandating sharing: journal policies in the natural sciences", *Science Communication*, Vol. 16 No. 4, pp. 403-431, available at: <https://doi.org/10.1177/1075547095016004003>
- Marcial, L.H. and Hemminger, B.M. (2010), "Scientific data repositories on the web: an initial survey", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 10, pp. 2029-2048, available at: <https://doi.org/10.1002/asi.21339>
- Murillo, A.P. (2014), "Examining data sharing and data reuse in the DataONE environment", *Proceedings of Association for Information Science and Technology (ASIS&T) Annual Meeting*, Vol. 51 No. 1, doi: [10.1002/meet.2014.14505101155](https://doi.org/10.1002/meet.2014.14505101155)
- National Institutes of Health (2003), "Final NIH statement on sharing research data", available at: <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>
- National Institutes of Health (2007), "NIH data sharing policy", available at: http://grants.nih.gov/grants/policy/data_sharing/
- National Science Foundation (2010), "Dissemination and sharing of research results", available at: www.nsf.gov/bfa/dias/policy/dmp.jsp
- Neuendorf, K.A. (2002), *The Content Analysis Guidebook*, Sage Publications, Thousand Oaks, Calif.
- Noor, M.A.F., Zimmerman, K.J. and Teeter, K.C. (2006), "Data sharing: how much doesn't get submitted to genBank?", *PLoS Biology*, Vol. 4 No. 7, p. e228, available at: <https://doi.org/10.1371/journal.pbio.0040228>
- Piwovar, H.A. (2011), "Who shares? Who doesn't? Factors associated with openly archiving raw research data", *PLoS One*, Vol. 6 No. 7, pp. e18657, 1-13, available at: <https://doi.org/10.1371/journal.pone.0018657>
- Piwovar, H.A. and Chapman, W.W. (2010), "Public sharing of research datasets: a pilot study of associations", *Journal of Informetrics*, Vol. 4 No. 2, pp. 148-156, available at: <https://doi.org/10.1016/j.joi.2009.11.010>
- Piwovar, H.A. and Vision, T.J. (2013), "Data reuse and the open data citation advantage", *PeerJ*, Vol. 1, p. e175, available at: <https://doi.org/10.7717/peerj.175>
- Sayogo, D.S. and Pardo, T. A. (2013), "Exploring the determinants of scientific data sharing: understanding the motivation to publish research data", *Government Information Quarterly*, Vol. 30 (Supplement 1), pp. S19-S13, available at: <https://doi.org/10.1016/j.giq.2012.06.011>
- Si, L., Li, Y., Zhuang, X., Xing, W., Hua, X., Li, X. and Xin, J. (2015), "An empirical study on the performance evaluation of scientific data sharing platforms in China", *Library Hi*

- Tech*, Vol. 33 No. 2, pp. 211-229, available at: <https://doi.org/10.1108/LHT-09-2014-0093>
- Sieber, J.E. (1988), "Data sharing: defining problems and seeking solutions", *Law and Human Behavior*, Vol. 12 No. 2, pp. 199-206.
- Someren, M.W., van Barnard, Y.F. and Sandberg, J. (1994), *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*, Academic Press, London; San Diego.
- Spurgin, K.M. and Wildemuth, B.M. (2009), "Content analysis", in Wildemuth, B.M. (Ed.), *Applications of Social Research Methods to Questions in Information and Library Science*, Libraries Unlimited, Westport Conn, pp. 189-198.
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E. and Frame, M. (2011), "Data sharing by scientists: practices and perceptions", *PLoS One*, Vol. 6 No. 6, pp. e21101. 1-21.
- Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B. and Dorsett, K. (2015), "Changes in data sharing and data reuse practices and perceptions among scientists worldwide", *Plos One*, Vol. 10 No. 8, p. e0134826, available at: <https://doi.org/10.1371/journal.pone.0134826>
- U.S. Geoscience Information Network (n.d.), "USGIN", available at: <http://usgin.org/> (accessed 8 January 2019).
- University of California Curation Center (2019), "DMPTool", available at: <https://dmpptool.org/> (accessed 19 February 2019).
- Yoon, A. (2017), "Role of communication in data reuse", *Proceedings of the Association for Information Science and Technology*, Vol. 54 No. 1, pp. 463-471.
- Yoon, A., Jeng, W., Curty, R. and Murillo, A. (2017), "In between data sharing and reuse: shareability, availability and reusability in diverse contexts", *Proceedings of the Association for Information Science and Technology*, Vol. 54 No. 1, pp. 606-609, available at: <https://doi.org/10.1002/pra2.2017.14505401085>
- Zimmerman, A.S. (2003), *Data sharing and secondary use of scientific data: Experiences of ecologists* (Order No. 3079559), ProQuest Dissertations & Theses Global, available at: <http://ulib.iupui.edu/cgi-bin/proxy.pl?url=http://search.proquest.com.proxy.ulib.uits.iu.edu/docview/287907131?accountid=7398>
- Zimmerman, A.S. (2008), "New knowledge from old data: the role of standards in the sharing and reuse of ecological data", *Science, Technology, & Human Values*, Vol. 33 No. 5, pp. 631-652.

Appendix 1: Post-result usefulness survey

On a scale of 1-5, with 1 being not useful and 5 being very useful, how would you rate this result in regards to assisting you in the ability to reuse the data?

Appendix 2: Post-search rank order survey

Please rank all results in the order of most useful to least useful in regards to assisting you in the ability to result the data.

Appendix 3: Post-experiment survey

Have you ever used DataONE Search system?

- Yes
- No
- Unsure

If yes, how often have you searched?

- Never
- Rarely
- Sometimes
- Quite Often
- Very Often

Open-ended questions

When looking at a search result, what information did you need to determine if the data is relevant?

In regards to the DataONE, what information inhibited your ability to determine data reusability?

In regards to the DataONE, what information facilitated your ability to determine data reusability?

When considering the DataONE, what information did you want about the data that the system did not provide?

Data Reuse Factors Questions

(IN GENERAL) When looking for data for reuse, what information do you need in order to determine data relevance and reusability?

Not at all important (1)

Very unimportant (2)

Somewhat unimportant (3)

Neither important nor unimportant (4)

Somewhat important (5)

Very important (6)

Extremely Important (7)

1. The data follows a specific metadata standard.
2. The data record contains information regarding provenance information.
3. The data record contains information regarding permissions and intellectual property rights.
4. The data record contains information regarding instrumentation.
5. The data record contains information regarding research methods.
6. The data record contains information regarding an attribute list.
7. Other: Please specify

Demographic questions

Sex

- Female
- Male
- Prefer not to answer
- I identify as/In another way (please specify if you wish): _____

Years of professional experience

- 0-5 years
- 6-10 years
- 11-20 years
- 20+ years

Educational background

- BA/BS
- MA/MS
- PhD
- Other

Area of Expertise (Please select all that apply)

- Ecology
- Geology
- Biology
- Atmospheric Science
- Environmental Science
- Hydrology
- Soil Science
- Chemistry

Corresponding author

Angela P. Murillo can be contacted at: apmurill@iu.edu

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgroupublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com